Supplementary material for Revisiting Fully Convolutional Geometric Features for Object 6D Pose Estimation

Jaime Corsetti
jaime.corsetti98@gmail.com

Davide Boscaini
dboscaini@fbk.eu

Fabio Poiesi
poiesi@fbk.eu

Fondazione Bruno Kessler, Italy

1. Introduction

We provide additional material in support of our main paper. The material is organised as follows:

- In Sec. 2 we show qualitative results on the LMO [1] and YCBV [6] datasets, divided into success (Figs. 1 & 3) and failure (Figs. 2 & 4) cases. We also report examples of how pose registration is affected by correct and incorrect detections. To assess the quality of the learnt features, we visualise the distances in the feature space between a point and all the other points (Figs. 5 & 6).
- In Sec. 3 we report an ablation study on one object of the YCBV dataset (Tab. 1).
- In Sec. 4 we report an additional ablation study on the choice of the t_{scale} hyperparameter.
- In Sec. 5 we show the detection metrics on the test set used in our results (Tabs. 3 & 4) and highlight problems we found in the ground-truth annotations (Fig. 7). We also report the percentage of cases in which the detector causes a failure or a success in the pose registration.

2. Additional qualitative results

In Figs. 1 & 3 we show examples of correct pose predictions on LMO [1] and YCBV [6], respectively. The top row shows the ground-truth poses, while the bottom row shows the results obtained with our experimental setting.

In Figs. 2 & 4 we show examples of wrong pose predictions on LMO [1] and YCBV [6], respectively. To highlight the detector contribution, we show the ground-truth poses (top row), our results with the object detection prior (middle row), and our results without the object detection prior (bottom row).

In LMO [1] we observe that the detector appears to be strongly influenced by the colour of the object, as it confuses the Can, Eggbox, and Glue objects which show similar colours (see Fig. 2(a-b-c)).

Something similar occurs in Fig. 2(d), where the pose prediction for Holepuncher is correct when the object detector prior is not used, and wrong when it is used. This is due to the colour similarity between Holepuncher and the toy car behind the Glue object.

In YCBV [6] the object detector handles important errors (Figs. 4(b-c)) and improves the pose prediction accuracy (Figs. 4(f-g-h)). We also show a failure case where a wrong detection causes an inaccurate pose prediction (the extra large clamp object in Fig. 4(e)).

To examine the point-level features learned by FCGF6D, we select pairs of point clouds and visualise the distance in the feature space of each point with respect to a reference point. Consider Figs. 5 & 6 for LMO and YCBV, respectively. Given a pair (O, S), respectively object and scene, we randomly select three points belonging to a correspondence on O, then we compute the distance in the feature space of each point in O and S from it. The distance is then normalised for better visualisation. We show the RGB point clouds on the left, and the input pairs with the feature distance on the right. The reference point is depicted as a red point on O.

In LMO [1] (Fig. 5), we can observe that the distance in the scene point clouds is small near the point corresponding to the reference one. We can also observe that the model can learn a certain degree of symmetry: in the second visualisation of the Can object, the distance in feature space of a point which is symmetric to the reference one is small.

In Fig. 6 we show the visualisation on YCBV [6]. In general the distribution of the distances appears to be noisier and less smooth than in LMO. We believe this to be due to the stronger sampling on S that we perform on YCBV, unlike the one done on LMO (20,000 points against 50,000 points). Another possible cause is that the RGB test images of YCBV are of a lower quality than the ones of LMO. As in the LMO visualisation, the corresponding point of the reference point in the scene has a small distance in the feature space. We can also observe the resulting distance in the case of similar objects: in the rightmost case of the extra large clamp, one of the points on a similar object (the large clamp) is similar to the reference one in the feature space. Because our features are learned to describe a local patch, their quality can be limited by the presence of similar

Table 1: Ablation study on the large marker object of YCBV. Performance are compared in terms of RRE [radians] and RTE [cm] errors (the lower the better), and FMR and ADD-S AUC (shortened to ADD) scores (the higher the better). Δ shows the improvement of each contribution in terms of ADD-S AUC with respect to the previous row.

	Improvements	RRE ↓	RTE ↓	FMR ↑	ADD ↑	Δ
	Baseline	2.0	4.6	0.00	77.2	–
Loss	+ $\tau_{NS} = 0.1D_S$	2.0	4.2	0.00	78.7	+1.5
	+ $\tau_{NS} = 0.1D_O$	2.0	4.3	0.00	78.3	−0.4
Arch.	+ Independent weights	2.0	4.1	0.00	79.4	+1.1
	+ Add RGB information	1.2	3.2	49.1	84.9	+5.5
Aug.	+ Color augmentation	1.2	3.3	50.0	84.6	−0.3
	+ Random erasing	1.2	3.1	53.4	85.2	+0.6
Optim.	+ SGD → Adam	0.0	0.4	100	97.5	+12.3
	+ Adam → AdamW	0.0	0.4	100	97.5	0
	+ Exp → Cosine	0.0	0.4	100	97.4	−0.1

geometric structures in other objects. Also in this case we can observe how symmetry influences the feature space, by considering the visualisation of the mug object. We can observe that, especially in the second and third pairs, the most similar points on S are the ones on the radial symmetry axis of O.

3. Ablation study on YCBV

In Tab. 1 we report the results of our ablation study on YCBV [6]. We choose the large marker object and train a single model on it for each modification we applied. Each model is trained for 20 epochs on the standard training set. For the computation of the Feature Matching Recall (FMR), we set the distance threshold $\tau_1 = 10$ voxels and the inlier ratio threshold $\tau_2 = 5\%$, to account for the different density of the scene point cloud in YCBV. All the other settings and parameters are the same as those in our ablation study on LMO [1] in the main paper.

We can observe that some changes do not increment performance, but instead cause a slight drop, in particular when adapting the safety threshold to the object dimension (third row, -0.4) and when colour augmentation is applied (sixth row, -0.3). These additions do not benefit this particular object, but are instead advantageous when averaging all the object in the dataset.

We can note that, as in the ablation study on the LMO dataset in the main paper, the most significant improvements in ADD-S AUC result from applying the safety threshold (+1.5), adding RGB information (+5.5), and using the Adam optimiser (+12.3).

4. Additional ablation study on LMO

We include in Tab. 2 an ablation study on the t_{scale} hyperparameter, which is used to set the radius of the ball volume in which negative mining around a certain point is not allowed. We train on the Can object of LMO using the standard setting, and varying only t_{scale}. We can observe that our choice of $t_{\text{scale}} = 0.1$ leads to the best result. When t_{scale} is increased, many candidate points are forbidden to be used as negatives, therefore decreasing the final performance. On the other hand, a lower t_{scale} implies negative pairs composed by points which are near in the 3D space. This reduces the performance, as similar points are forced to have different descriptors. Notably, the worst results is obtained when $t_{\text{scale}} = 0$, i.e. when no negative candidates are excluded.

Table 2: Ablation study on the Can object of LMO. Performance is shown in terms of ADD-0.1 (the higher the better) in function of the hyperparameter t_{scale}.

t_{scale}	0.0	0.01	0.05	**0.1**	0.5
ADD-0.1d	66.55	91.80	93.79	**93.95**	81.28

Table 3: Object detector performance on LMO [1]. On this dataset, failures and successes of $\Delta_{S\to F}$ and $\Delta_{F\to S}$ are measured in terms of ADD(S)-0.1d. Key: *: symmetric object.

Object	AP	AP_{50}	$\Delta_{S\to F}$	$\Delta_{F\to S}$
Ape	64.9	95.8	3.3	1.5
Can	82.8	99.3	1.9	0
Cat	69.6	90.8	1.5	0.1
Driller	80.0	98.7	0.8	0.4
Duck	77.9	98.2	0.7	2.4
Eggbox*	58.8	85.9	4.3	24.8
Glue*	55.1	87.3	5.1	2.3
Holepuncher	82.3	99.5	2.0	1.4
Average	71.4	94.4	1.2	2.7

5. Contribution of the object detection prior

We report in Tabs. 3 & 4 the performance of the YOLOv8 detector [4] on the test set of LMO [1] and YCBV [6], respectively. Following [5], we report the area-under-the-curve of the Average Precision, obtained by varying the IoU threshold with respect to a ground truth detection from 0.5 to 0.95 with a step of 0.05 (AP). We also report the recall on Average Precision, obtained with a fixed IoU threshold of 0.5 (AP_{50}). We also measure how the performances in ADD(S)-0.1d and ADD-S-0.1d change when using a detector.

Because LMO and YCBV use ADD(S)-0.1d and ADD-S AUC, respectively, to measure the performance change, we consider ADD(S)-0.1d for LMO and in ADD-S-0.1d for YCBV. As in the original definition [2], a success for the ADD-S-0.1d metric is obtained when the ADD-S error is

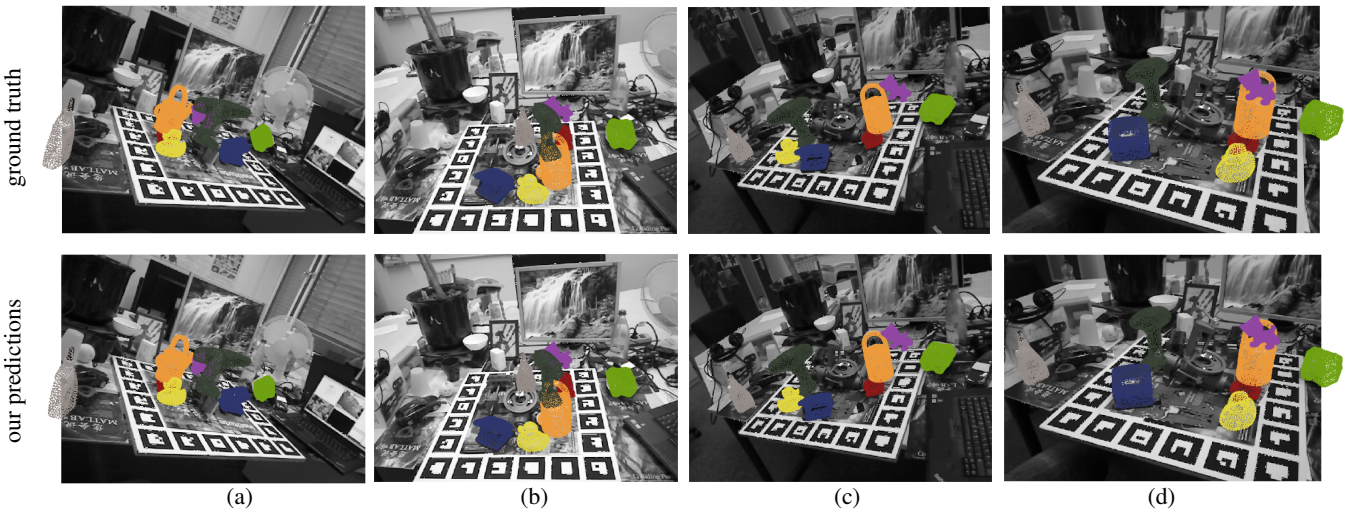

Figure 1: Example of success cases on the LMO dataset [1]. Colour key: ● Ape, ● Can, ● Cat, ● Drill, ● Duck, ● Eggbox, ● Glue, ● Holepuncher.

Figure 2: Example of failure cases on the LMO dataset [1]. Colour key: ● Ape, ● Can, ● Cat, ● Drill, ● Duck, ● Eggbox, ● Glue, ● Holepuncher.

below $0.1D_O$, where D_O is the object diameter, otherwise it is considered a failure. The same applies to ADD(S)-0.1d. Therefore, we define the following metrics:

- $\Delta_{S \to F}$: the percentage of object instances for which there is a success when not using a detector and a failure when using it.
- $\Delta_{F \to S}$: the percentage of object instances for which there is a failure when not using a detector and a success when using it.

In Tab. 3 we report the metrics on LMO [1]. We can observe that the detector performance is correlated with the size of the object: Ape, Cat and Duck are very small

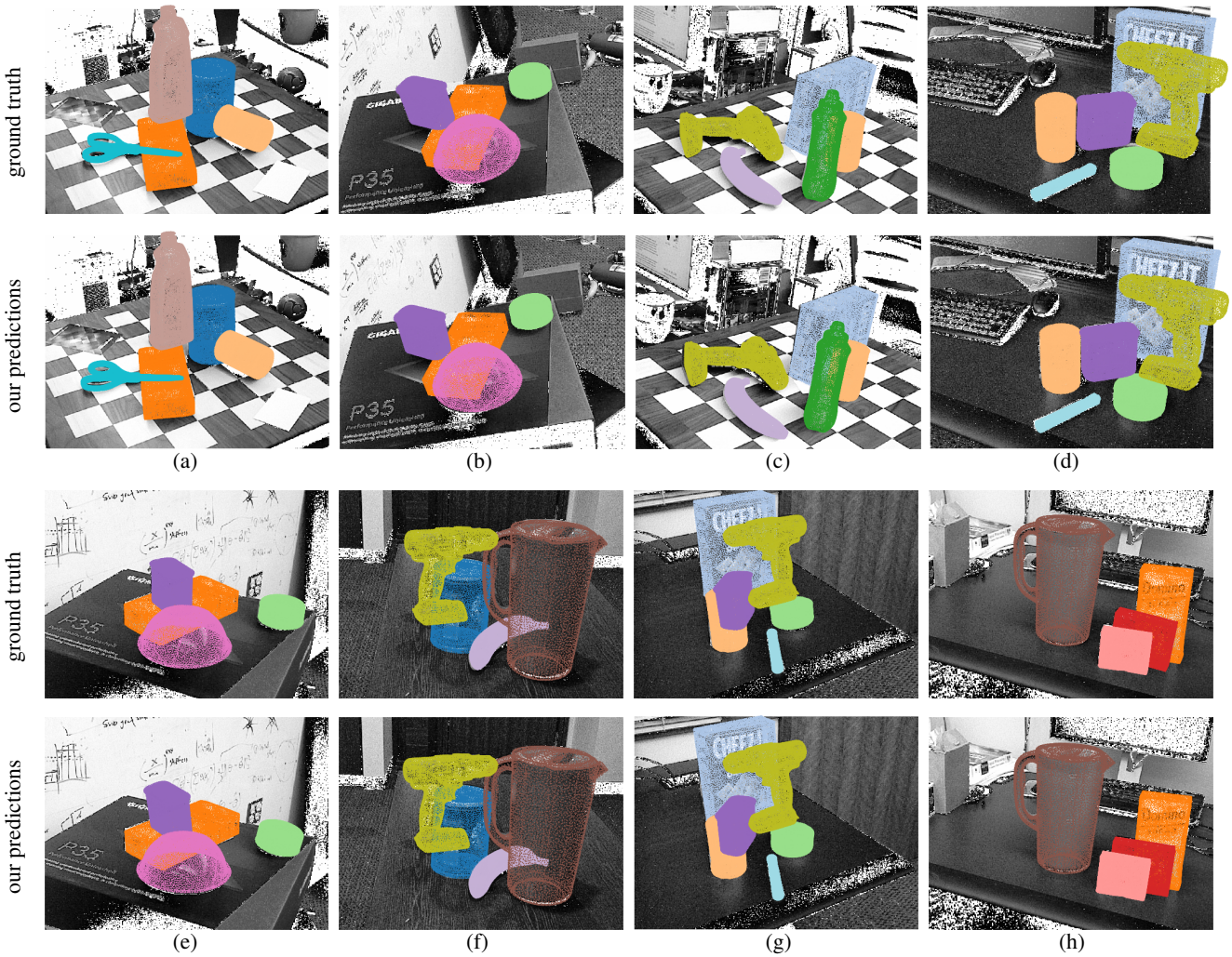

Figure 3: Example of success cases on the YCBV dataset [6]. Colour key: master chef can, cracker box, sugar box, tomato soup can, mustard bottle, tuna fish can, pudding box, gelatin box, potted meat can, banana, pitcher base, bleach cleanser, bowl, mug, power drill, wood block, scissors, large marker, large clamp, extra large clamp, foam brick.

and often occluded, and therefore perform worse than Can, Holepuncher, and Drill which are bigger. The lower performance on Glue and Eggbox objects occurs because they are often confused by the detector. An example of this can be observed in Fig. 2(b). This behaviour is also present with the Can object in Fig. 2(a). Despite these errors, we note that the Eggbox object greatly benefits from the detector (+24.8 in $\Delta_{F \to S}$), while for the other objects the performance gain is less significant.

In Tab. 4 we show the detection performances on YCBV [6]. Unlike LMO, the introduction of the object detector is beneficial for all the objects, as the $\Delta_{F \to S}$ is always higher then the $\Delta_{S \to F}$ (i.e. the detector solves more pose errors then it introduces for every object). We can note how the detector helps in solving the problem of object similarity: the large clamp and extra large clamp objects are amongst the ones that benefit the most from it. As an example, in Fig. 4(a) we show how the registration of the two objects (large clamp in dark blue, extra large clamp in dark green) differs depending on the use of the detector. With prior detection (second row) the two clamps are registered correctly. Without the detection (third row) the model registers both of them on the pose of the extra large clamp. We observe in Tab. 4 that the scissors object appears to be an outlier in the detector performance (AP_{50} of 27.6 against an average of 95.8). By examining the ground-truth detections we observed that they are noisy in the case of occlusion: sometimes the image portion where the object should be

Table 4: Object detector performance on YCBV [6]. On this dataset, failures and successes of $\Delta_{S\to F}$ and $\Delta_{F\to S}$ are measured in terms of ADD-S-0.1d. Key: *: symmetric object.

Object	AP	AP_{50}	$\Delta_{S\to F}$	$\Delta_{F\to S}$
master chef can	89.8	99.5	0.0	0.5
cracker box	91.5	99.5	0.1	0.3
sugar box	96.0	99.5	0.0	0.0
tomato soup can	85.3	98.8	0.0	5.1
mustard bottle	97.0	99.5	0.0	0.0
tuna fish can	88.5	99.5	1.9	18.3
pudding box	96.7	99.5	3.0	8.8
gelatin box	96.6	99.5	0.0	0.2
potted meat can	84.2	99.5	0.7	7.8
banana	90.9	99.5	0.0	0.1
pitcher base	99.4	99.5	0.0	0.0
bleach cleanser	88.3	99.4	0.2	1.6
bowl*	93.9	99.5	1.7	3.4
mug	91.5	99.5	0.0	0.4
power drill	95.4	99.5	0.0	0.0
wood block*	84.1	99.5	0.2	2.0
scissors	21.3	27.6	0.2	2.3
large maker	82.3	99.5	0.0	0.7
large clamp*	80.0	95.4	0.7	10.6
extra large camp*	83.0	97.8	1.5	20.1
foam brick*	86.3	99.5	0.2	4.0
Average	86.8	95.8	0.4	4.2

is also included in the bounding box, even if the object is not visible. See Fig. 7 for an example. Despite this, the scissors performance in pose estimation still benefits from the detector.

References

[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D object pose estimation using 3D object coordinates. In *ECCV*, 2014. 1, 2, 3, 7

[2] T. Hodan, J. Matas, and S. Obdrzalek. On evaluation of 6D object pose estimation. In *ECCV*, 2016. 2

[3] T. Hodan, M. Sundermeyer, B. Drost, Y. Labbe, E. Brachmann, F. Michel, C. Rother, and J. Matas. BOP challenge 2020 on 6D object localization. *ECCV Workshops*, 2020. 8

[4] G. Jocher, A. Chaurasia, and J. Qiu. YOLO by Ultralytics, 2023. 2, 8

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C.L. Zitnick. Microsoft COCO: Common Objects in COntext. In *ECCV*, 2014. 2

[6] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *RSS*, 2018. 1, 2, 4, 5, 6, 8

Figure 4: Example of failure cases on the YCBV dataset [6]. Colour key: master chef can, cracker box, sugar box, tomato soup can, mustard bottle, tuna fish can, pudding box, gelatin box, potted meat can, banana, pitcher base, bleach cleanser, bowl, mug, power drill, wood block, scissors, large marker, large clamp, extra large clamp, foam brick.

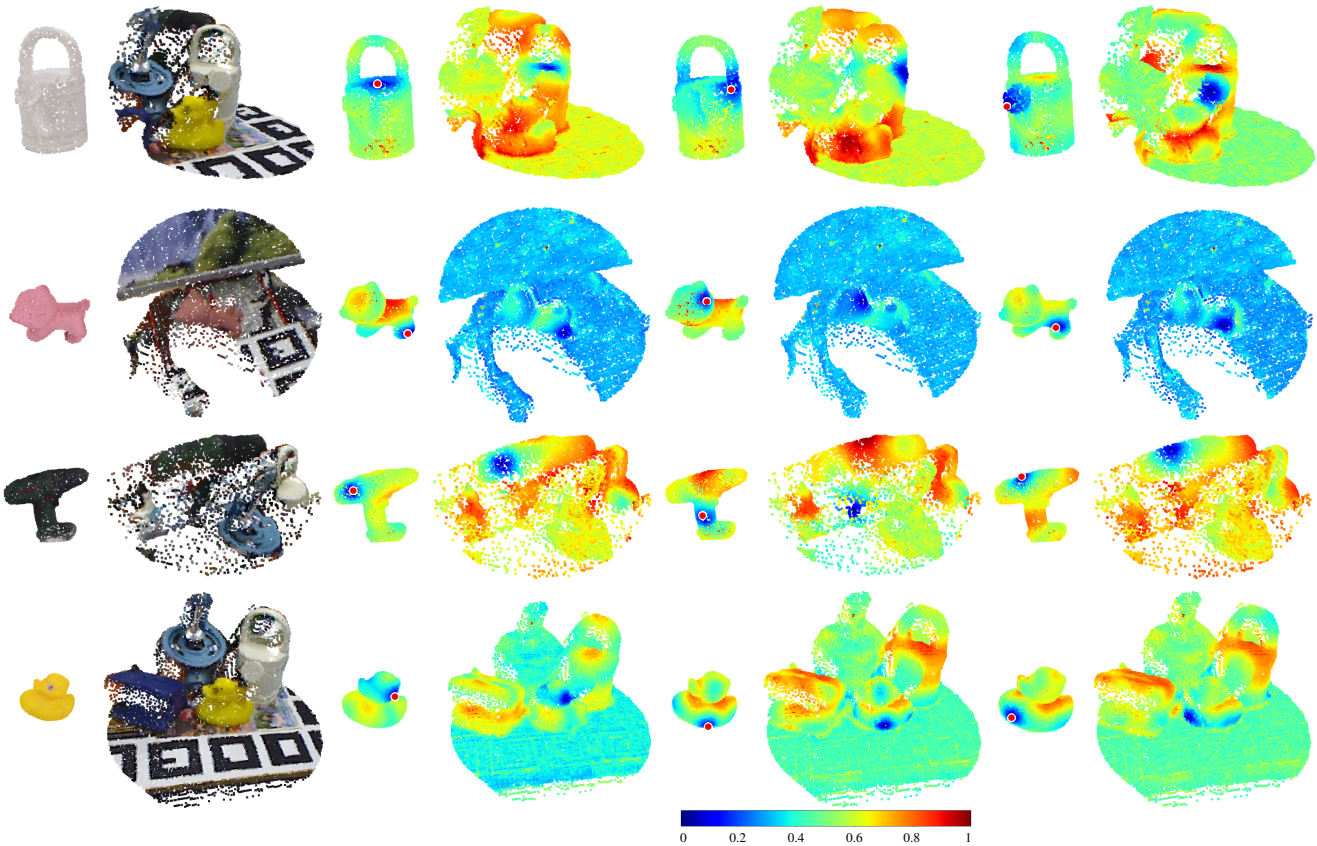

Figure 5: Qualitative evaluation of the features learned with our approach on the Can, Duck, Drill, and Cat objects of the LMO dataset [1]. We show the Euclidean distance between the features of a reference point (red point) on the object O and the features computed both on the rest of the points on O and on the scene S (cropped around the object of interest to facilitate visualisation). Cold and hot colours represent small and large distances, respectively. Ideal descriptors would produce a distance map with a sharp minimum at the corresponding point and no spurious local minima at other locations.

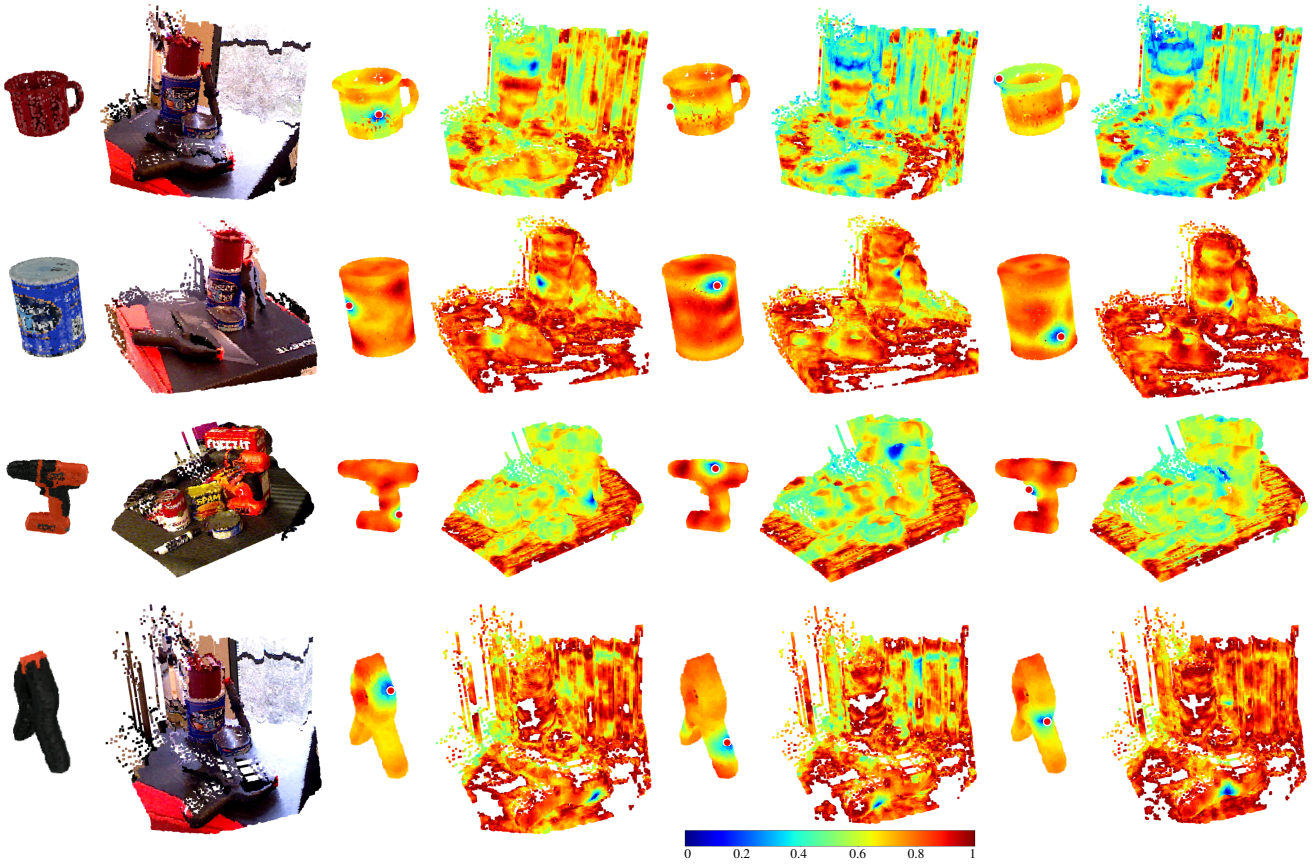

Figure 6: Qualitative evaluation of the features learned with our approach on the mug, master chef can, power drill, and extra large clamp objects of the YCBV dataset [6]. We show the Euclidean distance between the features of a reference point (red point) on the object O and the features computed both on the rest of the points on O and on the scene S (cropped around the object of interest to facilitate visualisation). Cold and hot colours represent small and large distances, respectively. Ideal descriptors would produce a distance map with a sharp minimum at the corresponding point and no spurious local minima at other locations.

Figure 7: Examples of noisy ground truth detections of the scissors object of YCBV [6] provided by the BOP challenge [3]. These annotations are used to train our YOLOv8 [4] object detector.